Lecture 1. Moment-generating functions.

Let Ω be a probability space and $X: \Omega \to \mathbb{R}$ a random variable.

<u>Def-n:</u> Let $f(x)$ be a pdf on $(X, \mathbb{R})$. Then

$$\mathbb{E}(X^n) = \int_{\mathbb{R}} x^n \cdot f(x)\, dx$$ is called the n-th moment of $f(X)$.

<u>Example</u>. (1) The first moment is $\mathbb{E}(X)$, the mean value.

(2) $n=2$. Using $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$ allows to find

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mathbb{E}^2(X).$$

<u>Idea:</u> suppose we have a sequence $\{a_n\}_{n \in \mathbb{N} \cup \{0\}}$ and want to 'store' it in a single f-n, i.e.

$$g(t) = \sum_{i=0}^{\infty} a_i t^i$$ with a dummy variable t.

<u>Example</u>. $a_i = \left(\frac{1}{2}\right)^i$ is a geometric progression.

$$g(t) = \sum_{i=0}^{\infty} \left(\frac{1}{2} \cdot t\right)^i = \frac{1}{1 - \frac{t}{2}}.$$

The f-n $g(t)$ is called a <u>generating f-n</u>.

Consider the f-n $M_X(t) := \mathbb{E}(e^{tX}) = 1 + \mathbb{E}(X) \cdot t + \frac{\mathbb{E}(X^2)}{2!} t^2 + \dots$ $\uparrow$ linearity of exp-n.

<u>Def-n</u>: $M_X(t)$ is called the <u>moment-generating f-n</u> (mgf) of a random variable X.

<u>Examples</u>:

(1). X is a Bernoulli random variable, i.e. X takes values 0 and 1, $P(X=1) = p$ and $P(X=0) = 1-p$.

$$M_X(t) = \mathbb{E}(e^{tX}) = (1-p) \cdot e^{t \cdot 0} + p \cdot e^{t \cdot 1} = 1 - p + pe^t.$$

(2) X is uniform on the interval $[0,2]$, i.e.

$$f_X(x) = \frac{1}{2-0} = \frac{1}{2}, \quad 0 \le x \le 2 \text{ and } f_X(x) = 0, \text{ otherwise.}$$

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_0^2 f_X(x) \cdot e^{tx} dx = \frac{1}{2}\int_0^2 e^{tx} dx = \frac{1}{2t} e^{tx}\Big|_0^2 =$$

$$= \frac{1}{2t}(e^{2t} - 1).$$

<u>Exercise</u>: Similarly, for X uniform on $[a,b]$ we get $M_X(t) = \dfrac{e^{tb} - e^{ta}}{t(b-a)}$.

(3) $X \sim N(0,1)$, i.e. normal distribution with mean $\mu = 0$ and st. dev. $\sigma = 1$.

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tx} e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx-\frac{x^2}{2}}\,dx = ?$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

To find this integral (the anti-der.), we follow the 2-step procedure:

① Complete the square in the power of exponent.

② Do a variable substitution to obtain a 'standard' Gaussian integral.

1. $tx - \frac{x^2}{2} = -\left(\frac{x}{\sqrt{2}}\right)^2 + 2\cdot\frac{x}{\sqrt{2}}\cdot\frac{t}{\sqrt{2}} - \frac{t^2}{2} + \frac{t^2}{2} = -\left(\frac{x-t}{\sqrt{2}}\right)^2 + \frac{t^2}{2}$

2. $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\left(\frac{x-t}{\sqrt{2}}\right)^2+\frac{t^2}{2}}\,dx = \frac{e^{t^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\left(\frac{x-t}{\sqrt{2}}\right)^2}\,dx =$

$\underset{u = x-t}{=} \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}}\,du = \frac{e^{t^2/2}}{\sqrt{2\pi}}\cdot\sqrt{2\pi} = e^{t^2/2}.$

③ Rmk: not all random variables have an mgf, since $M_X(t) = \mathbb{E}(e^{tX})$ may not be finite on any open interval containing 0.

Example. Let X have pdf $f_X(x) = \frac{1}{x^2}$, $x \geq 1$. Then

$$M_X(t) = \int_1^{\infty} \frac{1}{x^2} e^{tx}\,dx. \quad \left(\int_1^{\infty}\frac{1}{x^2}\,dx = -\frac{1}{x}\Big|_1^{\infty} = 1\right)$$

Notice that for any fixed $t > 0$, $\lim_{x\to\infty} \frac{e^{tx}}{x^2} = \infty$, so mgf does not exist.

Properties of mgf.

(1) $M_X(0)=1$

(2) Let X be a random variable with mgf $M_X(t)$, then the mgf of $aX+b$ is $M_{aX+b}(t)=e^{bt}M_X(at)$.

(3) Let $\{X_i\}_{i\in 1,\dots,n}$ be indep. random variables and $M_{X_i}(t)$ the corresponding mgfs. Then for $X=\sum_{i=1}^{n}a_iX_i$, we have

$$M_X(t)=M_{X_1}(a_1t)\dots M_{X_n}(a_nt),$$ here $a_1,\dots,a_n$ are constants.

(4) Let X and Y be two random variables with equal mgfs, i.e. $M_X(t)=M_Y(t)$. Then $F_X(x)=F_Y(x)$. (distrib. f-ns)

<u>Rmk</u>: This does not imply that $f_X(x)=f_Y(x)$. But this is true at 'almost all points'.

<u>Exercise</u>: verify (1)-(3), using properties of $\mathbb{E}$.

<u>Proof of (4)</u> in case X and Y are discrete random variables taking only finitely many values:

Let $S_X=\{x_1,\dots,x_n\}$ and $S_Y=\{y_1,\dots,y_m\}$ be the ~~prob.~~ values taken by ~~spaces of~~ X and Y. Similarly $S_{X\cup Y}=S_X\cup S_Y:=\{s_1,\dots,s_k\}$.

Then $M_X(t) = \sum_{i=1}^{k} p_X(s_i) \cdot e^{s_i t}$, where $p_X(s_i) = 0$, if $s_i \notin X$.

Similarly, $M_Y(t) = \sum_{i=1}^{k} p_Y(s_i) \cdot e^{s_i t}$.

As $M_X(t) = M_Y(t)$, we have $(*) \sum_{i=1}^{k} (p_X(s_i) - p_Y(s_i)) e^{s_i t} = 0$.

for t in some interval $(-\varepsilon; \varepsilon)$.

Since $(*)$ is true for inf. many values of t, we must have $p_X(s_i) = p_Y(s_i)$ for all i. $\square$.

<u>Rmk</u>: This is a very useful property, which allows to check if two distributions are identical.

<u>Example</u>. Let $X_1, \dots, X_n$ be Bernoulli random var-s, independent, with param. p.

Recall that binomial distr-n Y is the one with pdf

$p_Y(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$.

We will show that $X := X_1 + \dots + X_n \sim Y$ using property (4)

We checked that $M_{X_i}(t) = 1 - p + pe^t$. Using property (3), we have $M_X(t) = M_{X_1}(t) \dots M_{X_n}(t) = (1-p+pe^t)^n$.

Now let's compute $M_Y(t)$.

$$M_Y(t) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} e^{kt} = (pe^t + (1-p))^n.$$

$$\left((a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}, \text{ take } a = pe^t \text{ and } b = 1-p \right).$$

Other applications of mgfs.

(1) $\mathbb{E}(X) = M_X'(0)$

(2) $Var(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = M_X''(0) - (M_X'(0))^2$.

(3) $\mathbb{E}(X^n) = M_X^{(n)}(0)$.

Examples.

(1) X - Bernoulli random variable.

$M_X(t) = 1-p+pe^t$

$\mathbb{E}(X) = M_X'(t)\big|_{t=0} = pe^t\big|_{t=0} = p$.

$\mathbb{E}(X^2) = M_X''(t)\big|_{t=0} = pe^t\big|_{t=0} = p$.

$Var(X) = p - p^2 = p(1-p)$.

End of Lec. 1.

(2) $X \sim Norm(0,1)$, then $M_X(t) = e^{t^2/2}$.

$\mathbb{E}(X) = M_X'(t)\big|_{t=0} = te^{t^2/2}\big|_{t=0} = 0$.

$\mathbb{E}(X^2) = M_X''(t)\big|_{t=0} = (te^{t^2/2})'\big|_{t=0} = (e^{t^2/2} + t^2e^{t^2/2})\big|_{t=0} = 1$.

$Var(X) = 1 - 0 = 1$.

(3) $X \sim$ uniform on $(0,2)$, then $M_X(t) = \dfrac{e^{2t}-1}{2t}$

Let's find $\mathbb{E}(X) = M_X'(t)\big|_{t=0} = \dfrac{4te^{2t} - 2(e^{2t}-1)}{4t^2}\Big|_{t=0} = \dfrac{2te^{2t} - e^{2t}+1}{2t^2}\Big|_{t=0}$

This is where we need to use L'Hopital's rule!

$$\lim_{t\to 0} \frac{2te^{2t} - e^{2t}+1}{2t^2} = \lim_{t\to 0} \frac{(2te^{2t}-e^{2t}+1)'}{(2t^2)'} = \lim_{t\to 0} \frac{\cancel{2}e^{2t}(2t-1)+\cancel{2}e^{2t}}{\cancel{4}_2 t} \overset{\text{L'Hopital, again}}{=}$$

$$= \lim_{t\to 0} \frac{2e^{2t}(2t-1)+4e^{2t}}{2} = \frac{-1+2}{1} = 1.$$

Rmk: easier way, $\int_0^2 x \cdot \frac{1}{2}dx = 1$.

Example (HW exercise, p.9, #3).

Find the mgf, expected value and variance for the distribution $\underbrace{}_{\text{with pdf}}$ $f_X(x) = xe^{-x}$ for $x \geq 0$.

We check that $f_X(x)$, indeed, defines a pdf:

$$\int_0^\infty xe^{-x}dx \xlongequal{\text{by parts}} -xe^{-x}\Big|_0^\infty + \int_0^\infty e^{-x}dx = 0 - e^{-x}\Big|_0^\infty =$$

$$= 0-(-1) = 1 \checkmark.$$

$$M_X(t) = \int_0^\infty xe^{-x}\cdot e^{tx}dx = \int_0^\infty xe^{x(t-1)}dx.$$

Notice that the last integral diverge for $t \geq 1$, so we assume $t<1$ and set $z := x(t-1)$. Then $dz = (t-1)dx$. This substitution allows to write $\int_0^\infty xe^{x(t-1)}dx =$

$= \int_0^{-\infty} \frac{1}{(t-1)^2}e^z dz$ (notice that when $x \to \infty$, $z = x(t-1) \to -\infty$, since $t<1$).

$$= \frac{1}{(t-1)^2} e^z\Big|_0^{-\infty} = \frac{1}{(t-1)^2}.$$

$M_X'(t) = \frac{-2}{(t-1)^3} \Rightarrow \mathbb{E}(X) = M_X'(0) = \frac{-2}{-1} = 2.$

$M_X''(t) = \frac{6}{(t-1)^4} \Rightarrow \mathbb{E}(X^2) = M_X''(0) = \frac{6}{1} = 6.$

$\mathrm{Var}(X) = 6 - 2^2 = 2.$

<u>Rmk:</u> Recall the Central Limit Theorem (CLT): let $X_1, \dots, X_n$ be i.i.d. random variables with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then for $n \gg 0$, $\frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \to Z := N(0,1)$.

The CLT can be proved by showing that

$$\lim_{n \to \infty} M_n(t) = M_Z(t) = e^{t^2/2}, \text{ where } M_n(t) := M_{\frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma}}(t),$$

(see page 3 of the Handout File).

End of Lecture 1

The χ^2 distribution.

Consider $X_1, \dots, X_n$, i.i.d random variables with $X_i \sim N(0,1)$. Then the random variable $X := \sum_{i=1}^{n} X_i^2$ has a pdf $f_X(x) = \dfrac{x^{k/2-1} e^{-x/2}}{2^{k/2}\,\Gamma(k/2)}$, $x \geq 0$.

Rmk: the proof is straightforward (see p. 384 in [LM])

Here $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$.

Rmk: The χ^2 distribution is part of the family of Gamma distributions. These are distributions $\Gamma(\alpha,\beta)$ with pdfs $f(x,\alpha,\beta) = \dfrac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$.

Multinomial distr-n and goodness of fit.

The multinomial distr-n is a straightforward generalization of the binomial one.

Consider the situation where we have k possible outcomes $r_1, \dots, r_k$ with probabilities $p_1, \dots, p_k$, of course $\sum_{i=1}^{k} p_i = 1$.

If we repeat the experiment n times, the probability of getting the outcome r_1 on s_1 occassions, r_2 on s_2, ..., r_k exactly s_k times (here $\sum_{i=1}^{k} s_i = n$) is $p = \dfrac{n!}{s_1! \dots s_k!} p_1^{s_1} \dots p_k^{s_k}$.

Rmk: compare with the coefficient of $p_1^{\delta_1} \cdots p_k^{\delta_k}$ in the expression $(p_1 + \cdots + p_k)^n$.

Example. Consider an unfair die with $P(X=i) = \frac{2i-1}{36}$.

i	1	2	3	4	5	6
$P(X=i)$	$1/36$	$3/36$	$5/36$	$7/36$	$9/36$	$11/36$

Poll questions:

(1) The die is tossed 3 times. What is the probability that you get 6 on each occassion?

Answer: $\frac{\cancel{3}!}{\cancel{3}!} \cdot p_6^3 = \left(\frac{11}{36}\right)^3$

(2) The die is tossed 5 times. $P(S_1 = 1, S_2 = 2, S_3 = 2, S_{4,5,6} = 0) = ?$

Answer: $\frac{5!}{1!2!2!} p_1 \cdot p_2^2 \cdot p_3^2 = \frac{5!}{1!2!2!} \cdot \frac{1 \cdot 3^2 \cdot 5^2}{(36)^5} = 30 \cdot \frac{9 \cdot 25}{(36)^5} =$

$= \frac{6750}{36^5}$

Thm. Let $r_1, \ldots, r_k$ be the set of possible outcomes associated with each of n independent trials, where $P(r_i) = p_i$. Let $X_i =$ number of times r_i occurs.

(a) The random variable $X^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$ has approx. the χ^2 distr-n with $(k-1)$ degrees of freedom.

For an adequate approx-n, we need to have $np_i \geq 5$ for all i.

(b) Let $s_1, \dots, s_k$ be the observed frequences for the outcomes $r_1, \dots, r_k$, let $np_1, \dots, np_k$ be the corresponding expected values according to the null hypothesis. Then, at the α level of significance, H_0: $p_X(s) = p_0(s)$, i.e. $p_1 = p_{10}, p_2 = p_{20}, \dots, p_s = p_{s0}$ is rejected if

End of Lec.2 $d = \sum_{i=1}^{k} \frac{(s_i - np_{i0})^2}{np_{i0}} \geq \chi^2_{\alpha, k-1}$ (here $np_{i0} \geq 5$ $\forall i$)

<u>Example.</u> In Queensboro the local soccer team played 100 games last season with 60 wins, 25 losses and 15 draws. Test H_0: $p_{win} = 0.5$, $p_{Loss} = 0.3$ and $p_{Draw} = 0.2$ at the 5% level of significance.

outcome	Win	Loss	Draw
actual value	60	25	15
Expected value	50	30	20

$$d = \frac{(60-50)^2}{50} + \frac{(25-30)^2}{30} + \frac{(15-20)^2}{20} = 4.083.$$

$\chi^2_{0.05, 2} = 5.99$. Since $d < \chi^2_{0.05, 2}$, we accept H_0.

Two-way Tables.

Suppose we have a collection of observations consisting of measures on two variables. We would like to test if these measures are independent of each other.

Example.

Sex \ Handedness	Right handed	Left h.	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

To perform the test we follow the procedure.

Step 1. Compute the expected values $E_{ij} = \frac{\#\text{row}_i \cdot \#\text{col}_j}{\text{Total}}$

45,24	6,76	52
41,76	6,24	48
87	13	100

Rmk: If sex/handedness are independent then, say $P(L.h. \cap \text{Male}) = P(L.h.) \cdot P(\text{Male}) = \frac{13}{100} \cdot \frac{52}{100}$, so, we expect $\frac{13 \cdot 52}{100 \cdot 100} \cdot 100$ left handed males.

Step 2. Compute the test statistic

$$d = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$, where r = number of rows, c = number of columns

$$d = \frac{(43-45.24)^2}{45.24} + \frac{(9-6.76)^2}{6.76} + \frac{(44-41.76)^2}{41.76} + \frac{(4-6.24)^2}{6.24} =$$

$= 1.77$

Step 3. Find the critical value $\chi^2_{df,\alpha}$, where $df = (r-1)(c-1)$ and compare with d to decide whether we should accept H_0 at level of significance equal to alpha(α).

$$df = (2-1)(2-1) = 1.$$

$$\chi^2_{1,0.05} = 3.841$$

$d < \chi^2_{1,0.05} \Rightarrow$ accept $H_0 \Rightarrow$ independent.

End of Lec 3

Reminder on Significance Level.

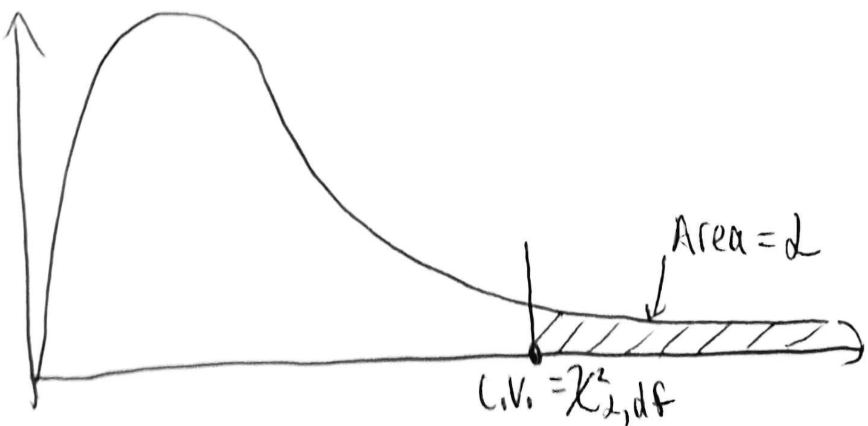
Def-n. Any function of the observed data whose numerical value dictates whether H_0 is accepted or not is called a test statistic.

Def-n The probability that the test statistic lies in the critical region (the set of values which result in rejection of H_0) is called the level of significance and is denoted by α.

Example. Goodness of fit.

We considered the random variable $X^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$, where p_i was the probability of the outcome r_i.

The hypothesis H_0 was $p_1 = p_{10}, \ldots, p_k = p_{k0}$ for a collection of probabilities $(p_{10}, \ldots, p_{k0})$. Then the pdf $f(X^2 | H_0)$ (given H_0) was approx. the $\underbrace{\chi^2_k}_{\text{one for}}$ distr-n with $k-1$ degrees of freedom.



If the test statistic $d > c.v.$, equivalently, $P(X^2 \geq c.v. | H_0) \hat{} \alpha$ (p-value test), we reject H_0.

Testing $H_0: \mu_X = \mu_Y$.

Let $x_1, \dots, x_n$ and $y_1, \dots, y_m$ be indep. random samples from normal distributions with means μ_X and μ_Y and same standard deviation σ.

Let $S_p^2 = \dfrac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$ be the pooled variance and

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{n+m-2}$$

$$T = \frac{\bar{x} - \bar{y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

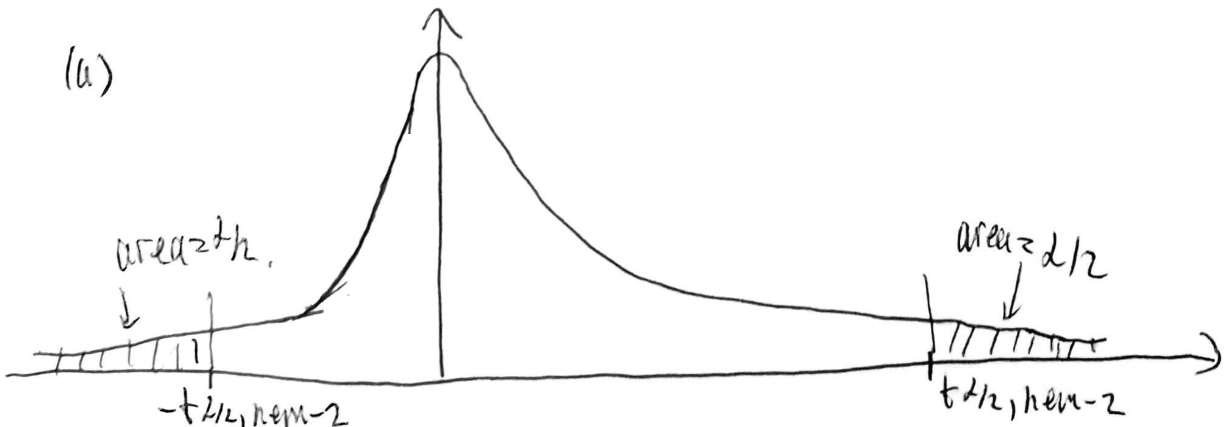
Then to test $H_0: \mu_X = \mu_Y$

(a) Against $H_1: \mu_X \neq \mu_Y$ at the α level of significance, reject H_0 if $|T| \geq t_{\alpha/2, n+m-2}$.

(b) $H_1: \mu_X > \mu_Y$, reject H_0 if $T \geq t_{\alpha, n+m-2}$

(c) $H_1: \mu_X < \mu_Y$, reject H_0 if $T \leq -t_{\alpha, n+m-2}$

(a)



Example. The University of Missouri gave a valodation test to entering students who had taken calculus in high school. The group of 93 students receiving no college credit had a mean score of 4.17 with a sample st. dev. of 3.70, while the group of 28 students who received credit from a high school dual enrollment class, had a mean score of 4.61 with sample st. dev. of 4.28. Is there a significant difference in these means at the $\alpha = 0.01$ level? (Assume the variances are equal)

Sol-n: $\bar{x} = 4.17$, $\bar{y} = 4.61$

$S_x = 3.7$, $S_y = 4.28$

$n = 93$ $\quad m = 28$

$$S_p^2 = \frac{92 \cdot 3.7^2 + 27 \cdot 4.28^2}{93+28-2} = 14.740$$

$$T = \frac{4.17 - 4.61}{S_p \sqrt{\frac{1}{93} + \frac{1}{28}}} = -0.532.$$

Crit. value $= t_{\frac{0.01}{2}, 119} \approx 2.617$. As $|T| < 2.617$, accept H_0, i.e. there is no significant difference.

ANOVA.

Analysis of var~~iations~~ variance, used to test if the means of three or more sample groups are the same.

Assumptions:

- The populations from which the samples are taken must be normally or approx. normally distributed
- The samples are independent
- The variances must be equal (σ^2)

Hypotheses.

H_0: all means are equal ($\mu_0 = \mu_1 = \dots = \mu_g$)
H_A: at least one mean is different.

Def-n: a balanced one-way ANOVA refers to the special case of one-way ANOVA in which all the numbers of observations in different groups are equal. An experimental layout with different numbers of observations is called unbalanced.

Rmk: 'one-way' stands for one independent variable (factor), i.e. the groups are 'parameterized' by a single variable.

Let X_{ij} denote the j^{th} measurement in the i^{th} population. Then $X_{ij} = \mu_i + \varepsilon_{ij}$, where ε_{ij} are i.i.d $N(0, \sigma^2)$ due to our assumptions.

The estimators of μ_i's are $\bar{X}_i$'s with $\bar{X}_i := \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}$.

Here $N_i :=$ number of elements in group i.

Define $\bar{X} := \frac{\sum_{i=1}^{g} N_i \bar{X}_i}{\sum_{i=1}^{g} N_i}$ to be the overall mean (experimental).

Set $SSG := \sum_{i=1}^{g} N_i (\overline{X}_i - \overline{X})^2$, i.e. 'sum of squares between groups'

<u>Thm.</u>: if H_0 is true, then $\frac{SSG}{\sigma^2}$ has a χ^2-distr-n with $g-1$ degrees of freedom.

Since we do not know σ, use the approximation

$$SSE := \sum_{i=1}^{g} (N_i - 1) S_i^2, \text{ where } S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \overline{X}_i)^2$$

Then SSE/σ^2 is also a χ^2 distr-n with $N-g$ degrees of freedom, $N = \sum_{i=1}^{g} N_i$.

<u>Thm.</u> (a) If H_0 is true then $F := \frac{SSG}{SSE} \cdot \frac{N-g}{g-1}$ is the Fisher distr-n $F_{g-1, N-g}$.

Recall that $F_{m,n}$ is the Fisher distribution with pdf

$$f_{F_{m,n}}(x) = \frac{\Gamma\left(\frac{m+n}{2}\right) m^{m/2} n^{n/2} x^{m/2 - 1}}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) (n+mx)^{\frac{m+n}{2}}}$$

(b) H_0 is rejected at α level of significance, if $F \geq F_{\alpha, g-1, N-g}$.

In practice, we will use the calculator for ANOVA (can be downloaded from TI website).

(see page 14 in 'Handout' for instructions)

Example. A hospital in Norwich is investigating a possible relationship between cigarette smoking and heart rates. Four factor levels ranging from Nonsmokers to Heavy Smokers were each represented by six subjects (see the Table below).

Are the differences among the groups statistically significant on the $\alpha = 5\%$ level?

Nonsmokers	Light Smokers	Moderate Smokers	Heavy Smokers
69	55	66	91
52	60	81	72
71	78	70	81
58	58	77	67
59	62	57	95
65	66	79	84

Let $\mu_1, \mu_2, \mu_3, \mu_4$ denote the true average heart rates in each of the groups. Then we need to test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$.

We find the mean estimators ('mean' stands for average, not 'cruel' here):

$\bar{X}_1 = 62.3$ and, hence, the overall experimental mean

$\bar{X}_2 = 63.2$ is $\bar{X} = \dfrac{6(62.3 + 63.2 + 71.7 + 81.7)}{24} = 69.7.$

$\bar{X}_3 = 71.7$

$\bar{X}_4 = 81.7$ (notice that $N_1 = N_2 = N_3 = N_4 = 6$ and $N = 24$).
$g = 4$

Next we find $SSG = 6\left((62.3-69.7)^2+(63.2-69.7)^2+(71.7-69.7)^2+(81.7-69.7)^2\right) = 1464.125.$

and $SSE = 5\cdot\frac{1}{5}\left(\sum_{i=1}^{4}\sum_{j=1}^{6}(x_{ij}-\bar{x}_i)^2\right) = \left[(69-62.3)^2+\ldots+(65-62.3)^2\right] + \ldots + \left[(91-81.7)^2+\ldots+(84-81.7)^2\right] = 1594.833.$

Finally, the test statistic is

$$F = \frac{SSG}{SSE}\cdot\frac{N-g}{g-1} = \frac{1464.125}{1594.833}\cdot\frac{24-4}{4-1} = 6.12.$$

On the other hand the critical value is

$F_{0.05,3,20} = 3.10$ (see page 24 of the 'Statistical Tables' file).

Since $F > F_{0.05,3,20}$, we reject H_0.

Using the calculator TI 84, we get the answer

Exercises (in preparation for Quiz 1)

p. 9, #1 (c).

Find the mgf for a random variable X with pdf

x	-1	0	1	10
$P(X=x)$	0.4	0.3	0.25	0.05

$$M_X(t) = \mathbb{E}(e^{tX}) = 0.4e^{-t} + \overset{0.3}{\overset{\shortparallel}{0.3e^{0\cdot t}}} + 0.25e^{t} + 0.05e^{10t}.$$

Using this result, find $\mathbb{E}(X)$ and $Var(X)$.

$$M_X'(t) = -0.4e^{-t} + 0.25e^{t} + 0.5e^{10t}.$$

$$M_X''(t) = 0.4e^{-t} + 0.25e^{t} + 5e^{10t}$$

$$\mathbb{E}(X) = M_X'(0) = -0.4 + 0.25 + 0.5 = 0.35.$$

$$\mathbb{E}(X^2) = M_X''(0) = 5.75.$$

$$Var(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = 5.75 - 0.1225 = 5.6275.$$

#1(d) $f_X(x) = 2x$ for $0 \le x \le 1$.

$$M_X(t) = \int_0^1 2x e^{tx}\,dx \overset{\text{by parts}}{=\!=} \frac{2x}{t}e^{tx}\Big|_0^1 - \int_0^1 \frac{2}{t}e^{tx}\,dx =$$

$$= \frac{2}{t}e^{t} - \frac{2}{t^2}e^{tx}\Big|_0^1 = \frac{2}{t}e^{t} - \left(\frac{2}{t^2}e^{t} - \frac{2}{t^2}\right) = \frac{2}{t}e^{t} - \frac{2}{t^2}(e^{t}-1).$$

#4. Let X be as in 1(d), find the mgf of $Y = 4X-2$.

Using property (2), i.e. $M_{aX+b}(t) = e^{bt}M_X(at)$, we get

$$M_Y(t) = e^{-2t}M_X(4t) = e^{-2t}\left(\frac{2}{4t}e^{4t} - \frac{2}{16t^2}(e^{4t}-1)\right) = \frac{1}{2t}e^{2t} - \frac{1}{8t^2}(e^{2t} - e^{-2t})$$

p. 12, #2.

Test $H_0: p_1 = 0.4, p_2 = 0.3, p_3 = 0.2, p_4 = 0.1$ at the 1% level, given the sample

outcome	1	2	3	4
Number of times	85	70	25	20

$k=4$

Total = 200

Expected values

1	2	3	4
80	60	40	20

$$d = \frac{(85-80)^2}{80} + \frac{(70-60)^2}{60} + \frac{(25-40)^2}{40} + \frac{(20-20)^2}{20} = 7.604$$

$\chi^2_{0.05,3} = 7.815$. $\chi^2_{0.01,3} = 11.345$

Since $d < \chi^2_{0.05,3}$, accept H_0 at the 5% level.

#5. A random sample of 300 households is chosen and people are asked where they live and their income (in $ thousand).

	≤50	50-100	≥100
Inside	131	74	37
Outside	38	15	5

Test at the 1% signif. level to see if the var-s are independent.

We will use the calculator TI 84.

1. Enter the table above in a matrix A.

$2^{nd} \to x^{-1}$ (MATRIX) $\to$ EDIT $\to$ 2×3 $\to$ ENTER (enter values).

2. $2^{nd} \to$ MODE (QUIT) $\to$ STAT $\to$ TESTS $\to$ χ^2 (enter A for observed values)

We get $\chi^2 = 2.913 < 9.21 = \chi^2_{0.01,2} \Rightarrow$ accept.