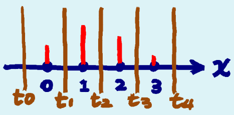
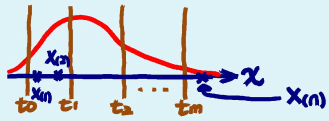
Remarks.

1. There is a distinction between $O_1, \ldots, O_m$ and $X_1, \ldots, X_n$ (especially for continuous case) empirical cdf (textbook, sec. 10.2) ⇕
   - $X_1, \ldots, X_n \Rightarrow X_{(1)}, \ldots, X_{(n)} \Rightarrow O_1, \ldots, O_m$
   - $O_1, \ldots, O_m \underset{\times}{\overset{\text{discrete (possible)}}{\Rightarrow}} X_{(1)}, \ldots, X_{(n)} \underset{\times \text{ if not i.i.d.}}{\overset{\text{if i.i.d.}}{\Rightarrow}} X_1, \ldots, X_n$ ← continuous

   If $X_1, \cdots, X_n$ are i.i.d, order statistics $X_{(1)}, \cdots, X_{(n)}$ are sufficient for any distribution

   $X_1, \cdots, X_n$: discrete r.v.'s





   $X_1, \cdots, X_n$: continuous r.v.'s

2. The MLE of θ based on $O_1, \ldots, O_m$ can be different from the MLE of θ based on $X_1, \ldots, X_n$.
3. Different choices of $(t_{i-1}, t_i]$, $i = 1, \ldots, m$, can cause different results. (**Note.** The choice should not depend heavily on observed data.)

   Note. In Ex.7.17, t_i's are not functions of data, i.e., t_i's are not statistics, not r.v.'s. ~~$t_i(X_1, \cdots, X_n)$~~

4. It is recommended that O_i, $E_i \geq 5$. ← a result guaranteed by large n ← $\boxed{\text{asymptotic property}}$

5/8

Example 7.19 (Hardy-Weinberg Equilibrium, TBp.343-344, or Ex.6.15, LN, Ch8, p.24)

- $n = 1029$, the cell probabilities are $\underbrace{(1-\theta)^2}_{AA}$, $\underbrace{2\theta(1-\theta)}_{Aa}$, $\underbrace{\theta^2}_{aa}$ under the Hardy-Weinberg Equilibrium model and the MLE of θ is $\hat{\theta} = 0.4247$.

$X_1, \cdots, X_{1029} \overset{\text{i.i.d.}}{\sim}$ multinomial$(1, p_1, p_2, p_3)$
$X_{(1)}, \cdots, X_{(1029)}$
$O_1, O_2, O_3 \sim$ multinomial$(1029, p_1, p_2, p_3)$

	Blood Type		
	M	MN	N
O_i	342	500	187
E_i	340.6	502.8	185.6

→ intuitive conclusion?

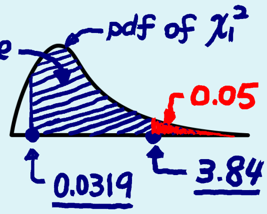
how to get this? e.g. $340.6 = n p_1(\hat{\theta}) = 1029 \cdot (1 - 0.4247)^2$

- Consider the test:

  Ω_0: H_0: $(p_1(\theta), p_2(\theta), p_3(\theta))$ are specified by the Hardy-Weihberg model

  H_A: (p_1, p_2, p_3) do not have that specified form → $\Omega: p_1 + p_2 + p_3 = 1$

- Pearson's chi-square test: $\boxed{\dim(\Omega_0) = 1,\ \dim(\Omega) = 2}$
  1. Pearson's chi-square test statistic is
  $$X^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{D} \chi_1^2 \text{ under } H_0.$$
     (m) $\quad \because n = 1029 \quad$ asymptotic null distribution $\quad$ $\dim(\Omega) - \dim(\Omega_0)$ $\quad$ meaning?



  2. Set $\alpha = 0.05$. Thus, reject H_0 if the value of X^2 statistic exceeds 3.84, the 95%-quantile of the χ_1^2 distribution. ← critical value
  3. Since
  $$X^2 = \frac{(\overbrace{342 - 340.6}^{1.4})^2}{340.6} + \frac{(500 - 502.8)^2}{502.8} + \frac{(\overbrace{187 - 185.6}^{1.4})^2}{185.6} \doteq 0.0319,$$
     H_0 is not rejected.

2. Why $\alpha = 0.05$? Will conclusion be different if we choose other α?
The p-value is more useful:

$$p\text{-value} = P_{H_0}(X^2 > \underbrace{0.0319}_{X^2_{obs}}) = P(\chi^2_1 \geq 0.0319) = 0.86.$$

more extreme

H_0 If the null model were correct, deviations this large or larger would occur 86% of the time. Thus, the data give us no reason to doubt the null model.

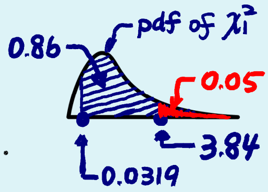
as extreme or more extreme

LNp.42

Why?

- Likelihood ratio statistic is

$$-2\log\Lambda = 2\sum_{i=1}^{3} O_i \log\left(\frac{O_i}{E_i}\right) = 0.032 \quad (\approx 0.0319 = X^2).$$



The two tests leads to the same conclusion. meaning?

Note: $\Lambda = \exp(0.032/(-2)) = 0.98 \approx 1$ $(0 \leq \Lambda \leq 1)$. Hardy-Weinberg model is almost as likely as the most general possible model.

Example 7.20 (Bacterial Clumps, TBp. 344-345)

- In testing milk for bacterial contamination, 0.01mL of milk is spread over an area of 1cm^2. 400 counts of bacterial clumps:

sample size

$X_1, \cdots, X_{400}$

$X_1, \cdots, X_{400}$ i.i.d. $P(\lambda)$

$O_1, \cdots, O_m$

number per 1cm^2	0	1	2	$\cdots$	9	10	19
Frequency	56	104	80	$\cdots$	3	2	1

histogram

possible statistical modeling
for $X_1, \cdots, X_{400}$?
for $O_1, \cdots, O_m$?

- H_0: The data are from Poisson $P(\lambda)$

- MLE for the λ of Poisson model (H_0) is

$$\overline{X} = \hat{\lambda} = \frac{0 \times 56 + 1 \times 104 + \cdots + 19 \times 1}{400} = 2.44,$$

Note. The MLE $\hat{\lambda}$ is the MLE based on $X_1, \cdots, X_{400}$, not the MLE based on $O_1, \cdots, O_m$, But, in the case, their MLEs are identical.

giving the expected frequencies E_i in the following table.

number per 1cm^2	0	1	2	$\cdots$	6	≥ 7
O_i (Obs. freq.)	56	104	80	$\cdots$	9	20
E_i (Exp. freq.)	34.9	85.1	103.8	$\cdots$	10.2	5.0
Component of χ^2	12.8	4.2	5.5	$\cdots$	0.14	45.0

Why?

23.8

15

$\frac{(O_i - E_i)^2}{E_i}$

how to get it?

$O_1, \cdots, O_8 \sim$ multinomial$(400, p_1, \cdots, p_8)$

H_0: $p_1(\lambda), \cdots, p_8(\lambda)$

critical value

- The chi-square statistic is $X^2 = 75.6 > 18.55 = \chi^2_6(0.005)$. So, p-value < 0.005 and the goodness of fit test rejects the Poisson model (H_0).

$dim(\Omega) = 7$, $dim(\Omega_0) = 1$

Why?

How assumptions in Poisson violated?

1. bacteria held by surface tension on lower surface of the drop may adhere to the glass slide on contact.
2. film not of uniform thickness.

New statistical modeling: Poisson → Negative binomial (LN. CH8. p.68)

Example 7.21 (Fisher's reexamination of Mendel's data, TBp. 345-346)

- Mendel crossed 556 smooth, yellow male peas with wrinkled, green female peas.

n

AA Aa aA BB Bb bB aa bb

$X_1, \cdots, X_{556} \overset{\text{i.i.d.}}{\sim} \text{multinomial}(1, P_1, P_2, P_3, P_4)$

	Type	Observed Count		Expected Count		Probability
$\{AA, Aa, aA\} \times \{BB, Bb, bB\}$	Smooth yellow	O_1 315	$O_1, \cdots, O_4 \sim \text{multinomial}(556, P_1, \cdots, P_4)$	E_1 312.75	How to get them?	P_{10} 9/16
$\{AA, Aa, aA\} \times \{bb\}$	Smooth green	O_2 108		E_2 104.25		P_{20} 3/16
$\{aa\} \times \{BB, Bb, bB\}$	Wrinkled yellow	O_3 102		E_3 104.25		P_{30} 3/16
$\{aa\} \times \{bb\}$	Wrinkled green	O_4 31		E_4 34.75		P_{40} 1/16

- For the data, Ω_0: $P_1 = \frac{9}{16}, P_2 = \frac{3}{16}, P_3 = \frac{3}{16}, P_4 = \frac{1}{16}$, $\Omega = \{(P_1, \cdots, P_4) \mid P_1 + P_2 + P_3 + P_4 = 1\}$

  $-2 \log \Lambda = 2 \sum_{i=1}^{4} O_i \log(O_i/E_i) = 0.618$, $\dim(\Omega) = 3$, $\dim(\Omega_0) = 0$

  $\Lambda = \exp(-0.618/2) = 0.73$, the p-value is slightly ← asymptotic null distribution: χ^2_3 ($n = 556$)

  less than 0.9. (Peason's statistic is $X^2 = 0.604$.)
- Fisher pooled the results of all of Mendel's experiments:
  - Two independent experiments give chi-square statistic T_1, T_2 (indep; $n_1 = 556$, n_2; from different datasets)
    with p and r degrees of freedom under H_0.
  - Under H_0, $T_1 + T_2 \sim \chi^2_{p+r}$. ← $p = r = 3$ in the case.
  - Adding all the chi-square statistics for all the independent experiments gives p-value=0.99996! ← too good to be true
- The best explanation is perhaps that Mendel continued experimenting until the results looked good. The statistical analysis here assume n is fixed before data are collected. Then, n is a random variable

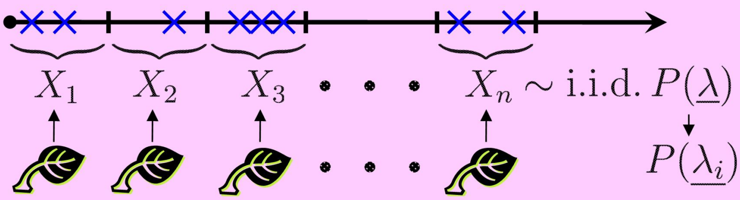
- Dorfman (1978) studied the goodness of fit of the intelligence scores of fathers and sons to a normal distribution (H_0) using Pearson's chi-square test. The p-values were greater than $1 - 10^{-7}$ and $1 - 10^{-6}$, respectively. → too good to be true?

❖ **Reading**: textbook, 9.5

- Application of GLR test II --- Poisson dispersion test

Question 7.16

- Recall the insect counts example (Ex. 6.31, LN, Ch8, p.68), the Poisson model did not fit well.
- Poisson model assumptions:



  1. The rate is constant. (一次單位時間)
  2. Counts in one interval are independent of counts in disjoint intervals.
- For counts of insects on leaves, some assumptions may be violated, e.g.,
  - Leaves are of different sizes and occur at various locations on different plants. Hence, 1. may fail.
  - If the insects hatched from eggs that were deposited in groups, there may be clustering of the insects. Then, 2. may fail.
- How to examine whether the rate is a constant for the insect data? → But, still assume $X_1, \cdots, X_n$ are independent Poisson.

Example 7.22 (GLR test for Poisson dispersion, TBp. 347-348)

cf. Ex.7-20 (LNp.45)

- statistical modeling for solving the question:
  - $X_i \sim P(\lambda_i), i = 1, \ldots, n$. & $X_1, \cdots, X_n$ are independet → joint pmf $\prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}$

    $$\Omega = \{(\lambda_1, \ldots, \lambda_n) : \lambda_i > 0, i = 1, \ldots, n\} \quad \Rightarrow \dim(\Omega) = n.$$
  - Null hypothesis H_0: the counts are Poisson with common parameter λ.

    $X_1, \cdots, X_n$ i.i.d. $P(\lambda)$

    $$\Omega_0 = \{(\lambda_1, \ldots, \lambda_n) : \lambda_1 = \cdots = \lambda_n \equiv \lambda\} \quad \Rightarrow \dim(\Omega_0) = 1.$$
  - Alternative hypothesis H_A: the counts are Poisson with different rates $\lambda_1, \lambda_2, \ldots, \lambda_n$, i.e., $\Omega_A = \Omega \backslash \Omega_0$.
- Under Ω_0, MLE of λ is $\hat{\lambda} = \overline{X}$.
- Under Ω, MLE of λ_i's are X_i's, denoted by $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_n$.
- Thus the likelihood ratio is

$O_i \leftrightarrow X_i$, $E_i \leftrightarrow \overline{X}$ cf.

test statistic in Ex.7.17 (LNp.40)

$$\Lambda = \frac{\prod_{i=1}^{n} \hat{\lambda}^{x_i} e^{-\hat{\lambda}} / \cancel{x_i!}}{\prod_{i=1}^{n} \tilde{\lambda}_i^{x_i} e^{-\tilde{\lambda}_i} / \cancel{x_i!}} = \prod_{i=1}^{n} \left(\frac{\overline{x}}{x_i}\right)^{x_i} e^{x_i - \overline{x}}$$

test statistic in Ex.7.18 (LNp 41) cf.

$$-2\log\Lambda = 2\sum_{i=1}^{n}\left[x_i \log\left(\frac{x_i}{\overline{x}}\right) + (x_i - \overline{x})\right] = 2\sum_{i=1}^{n} x_i \log\left(\frac{x_i}{\overline{x}}\right)$$

check the proof in LNp 42 → Under H_0

$$\approx \frac{1}{\overline{x}}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n\hat{\sigma}^2}{\overline{x}}$$

← (reasonable?) $\hat{\sigma}^2$: sample variance

dispersion index, if >> 1

- $\hat{\sigma}^2/\overline{x}$: measure of clustering ← check the graph in LNp 48
- Null Poisson model: variance=mean.
- Alternative Poisson model: variance>mean.
- (asymptotic) null distribution: χ^2_{n-1} ← $\dim(\Omega) - \dim(\Omega_0)$ check LNp.39

• $Y \sim P(\lambda)$ ⇒ $E(Y) = \lambda$, $Var(Y) = \lambda$
• If $Y_1, \cdots, Y_n$ i.i.d. $P(\lambda)$: $\overline{Y} = \widehat{E(Y_i)}$, $\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n} = \widehat{Var(Y_i)}$

⊙ Poisson dispersion test (for goodness-of-fit)

use original data $X_1, \cdots, X_n$, not $O_1, \cdots, O_m$

- has high power against alternative that are overdispersed relative to Poisson, ← e.g. $X_1, \cdots, X_n$ i.i.d. negative binomial (LN, CH8, p.65~66)

  i.e. $Var(Y) >> E(Y)$

⊖ often used when there is not enough data to be accumulated into several cells. ← $X_1, \cdots, X_n \rightarrow O_1, \cdots, O_m$ (better to have $O_i \geq 5$)

Example 7.23 (Asbestos Fibers, Poisson dispersion test, TBp. 348)

- For the data in Ex. 6.4, LN, Ch8, p.9, $n = 23$ ← not large

$n\hat{\sigma}^2/\overline{x} = \frac{1}{\overline{x}}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 26.56$; $-2\log\Lambda = 2\sum_{i=1}^{n} x_i \log\left(\frac{x_i}{\bar{x}}\right) = 27.11$

- $\dim(\Omega) - \dim(\Omega_0) = 23 - 1 = 22$ → asymptotic null dist: χ^2_{22} ← questionable
- p-value for 27.11 is 0.21. So there is not enough evidence against the null hypothesis.
- **Note.** Sample size 23 is small and the test may have low power.

Example 7.24 (Bacterial Clumps, Poisson dispersion test, TBp. 348-349)

- For the data in Ex.7.20, LNp.45, $n=400$.
$$\bar{x}=2.44, \qquad \hat{\sigma}^2=4.59 \quad \Rightarrow \quad \frac{n\hat{\sigma}^2}{\bar{x}}=\frac{400\times 4.59}{2.44}=752.7$$
- The p-value is: S — RR of test 1, RR of test 2

$D \to N(0,1)$

$$p\text{-value} = P\left(\frac{n\hat{\sigma}^2}{\bar{X}}\geq 752.7 \,\Big|\, H_0\right) = P\left(\frac{\frac{n\hat{\sigma}^2}{\bar{X}}-399}{\sqrt{2\times 399}}\geq \frac{752.7-399}{\sqrt{2\times 399}}\right)$$
$$\approx 1-\Phi(12.5)\approx 0 \qquad (\text{normal approximation to } \chi^2_{m=399})$$

compare it with the p-value in Ex.7.20 (LNp.46)

$\dim(\Omega)=400$, $\dim(\Omega_0)=1$

- Thus, there is almost no doubt that the Poisson distribution fails to fit the data.

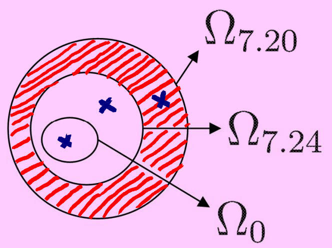
$\chi^2_m \sim Y = Y_1+\cdots+Y_m \xrightarrow{CLT} \text{Normal}$, $Y_i \sim \chi^2_1$

Question 7.17

- Compare Ex. 7.20 and Ex. 7.24. They test the same null hypothesis H_0. Why are the test statistics in the two examples different?
  - In Ex. 7.20, ← goodness-of-fit test $O_i \approx E_i$?
    $\Omega_{7.20}=\{X_i \text{ can be any discrete r.v.'s}\}$
  - In Ex. 7.24, ← Poisson dispersion test   variance $\approx$ mean?
    $\Omega_{7.24}=\{X_i \sim P(\lambda_i), i=1,\ldots,n.\}$

$X_1,\cdots,X_{400}$ i.i.d. $P(\lambda)$

specified



- Is it appropriate to use the test statistic in Ex. 7.20 to test the H_0 and H_A in Ex. 7.24? How is the opposite?
- For same data set, which of the tests in the two example would be expected to have smaller p-value? Why? ← (i) $\theta \in \Omega_{7.24}\setminus\Omega_0$ (ii) $\theta\in\Omega_{7.20}\setminus\Omega_{7.24}$

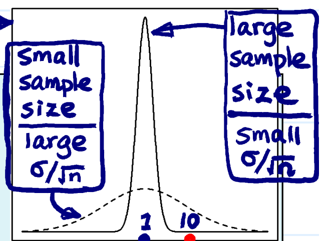
Note. If one has a specific alternative hypothesis in mind, better power can be obtained by developing a test against that alternative rather than against a more general alternative.

❖ **Reading**: textbook, 9.6

5/13

- Some concerns about hypothesis testing

pdf of $\bar{Y}_n$



- Question: Suppose modeling is correct. For
$$H_0\colon \theta=\theta_0 \quad \text{vs.} \quad H_A\colon \theta\neq\theta_0$$
when H_0 is not rejected, does it mean we accept $\theta=\theta_0$?

e.g.: $Y_1,\ldots,Y_n \sim N(\mu,\sigma^2)$, σ known,
$\mu\approx 0$, but not zero, reject H_0 if

$H_0: \mu=0$
$H_A: \mu\neq 0$

$\bar{Y}_{10}=10$, not reject
$\bar{Y}_{10000}=1$, reject

0 μ $\hat{\mu}=\bar{Y}_n$ $c\cdot\frac{\sigma}{\sqrt{n}}$

$$\left|\frac{\bar{Y}-0}{\sigma/\sqrt{n}}\right|\geq c \Leftrightarrow |\bar{Y}|>c\,\frac{\sigma}{\sqrt{n}}=c\sqrt{Var(\bar{Y})}.$$

LNp.28

$\frac{\sigma}{\sqrt{n}}\downarrow$ when $n\uparrow$
$\frac{\sigma}{\sqrt{n}}\downarrow$ when $\sigma\downarrow$

Consider the two cases:
(i) $n=10$, and (ii) $n=10000$.

$\bar{Y}=10$ not reject $\bar{Y}=1$ reject